August 2022

AACI Commentary

Democratizing Artificial Intelligence

Bringing Automated Machine Learning Technologies to Basic, Translational, and Clinical Cancer Researchers

By Dan Theodorescu, MD, PhD, and Jason H. Moore, PhD

Dan Theodorescu, MD, PhD, is the director of Cedars-Sinai Cancer Center.

Jason H. Moore, PhD, is chair of the Department of Computational Biomedicine at Cedars-Sinai Medical Center.

Commentary Overview

- Machine learning, a subdiscipline of artificial intelligence focused on data analytics, has played a key role in cancer research and care. This is due, in part, to the complexity of the disease and the availability of big data from technologies such as genomics and imaging.
- Cancer research can be greatly accelerated by removing the accessibility/learning curve bottleneck to the machine learning toolbox.
- Machine learning should be democratized so everyone can benefit from it with minimal time investment, without relying on a computational or statistical expert, and without a large budget.

Artificial intelligence (AI) originated in the 1950s after the first computers were built and deployed during World War II. Prior to this, AI was the subject of science fiction and referred to by different names, such as cybernetics. The use of the phrase "artificial intelligence" was solidified by leaders in the field at a Dartmouth College workshop in the summer of 1956. AI took off in the 1960s and 1970s as computers became more commonplace and programming languages such as FORTRAN and LISP matured.

The 1970s saw some of the first AI methods and software in the health care space. These included the MYCIN expert system for prescribing antibiotics for ICU patients that yielded more

accurate results compared to clinicians[1-4]. Although there were some early successes with AI in healthcare, there were also numerous failures that helped usher in an era referred to as the "AI winter," which lasted from the 1980s through the early 2000s[5-7]. This was a period characterized by dampened enthusiasm for AI, resulting from overhyping and under-delivering on the promises of the technology. This all changed in 2011 when IBM Watson wowed the tech industry and a corner of U.S. pop culture with its win against two of the greatest champions on *Jeopardy!*[8-9]. This demonstrated convincingly that AI was ready to take on humans at some tasks, including some medical decision making[10].

Machine learning, a subdiscipline of AI focused on data analytics, has played a prominent role in cancer research and care. In part, this is due to the complexity of the disease and the availability of big data from technologies, such as genomics and imaging. Applications include predicting regulatory elements in DNA sequences, predicting disease risk in populations, and diagnosing cancer from pathology and radiology images, as well as modeling and predicting physiologic and biologic behaviors or systems biology[11-12]. Despite the importance of machine learning in this domain, the methods and software are only accessible to those with training and expertise in the data, computer, and statistical sciences. This is because the algorithms and methods are not intuitive, are quite complex, and produce results that are not easily interpreted or explained.

Accelerating Cancer Research With Machine Learning

It is our working hypothesis that basic, translational, and clinical cancer research can be greatly accelerated by removing the accessibility/learning curve bottleneck to the machine learning toolbox. Machine learning should be democratized such that everyone who wants to benefit from it can do so with minimal time investment, and without relying on a computational or statistical expert who might have limited time or availability, and without the need for a large budget. The lack of accessible machine learning methods and software is reminiscent of the early days of statistics when statistical software was difficult to use and primarily accessible to statisticians. This changed in the 1990s as graphical user interfaces matured, making it much easier for anyone to load a data set, select a statistical method, and visualize the result. Additional features in the software programs—such as data quality control, the testing of assumptions and built-in educational resources and help in the software programs in test selection and utility—meant that anyone could perform a competent statistical analysis. It can certainly be argued that consulting an expert is a good idea. However, empowering biologists and clinicians to perform their own statistical analyses can increase the rate of scientific discovery and advance the statistical knowledge and experience of the user. Perhaps equally important, it drives cultural appreciation of the fact that today, and in the future, understanding biology and medicine requires increasingly quantitative tools and analysis to fully leverage the massive data and promise of genomic science.

A central challenge of machine learning is choosing algorithms and deciding how to tune the many hyperparameters that govern how they work. First, it might be of interest to select a subset of variables, or features, from thousands or millions that might be available. There are dozens of feature selection algorithms, and each looks at the data differently. Second, the features might need to be transformed prior to analysis. There are dozens of methods for transforming features, but one of the most well known is log transformation of data to facilitate and improve linear regression[13]. Third, the features might need to be combined and recoded based on their biological or mathematical relationships. This is called feature engineering and there are dozens of methods that could be used. Finally, there are many algorithms for building predictive models. Each algorithm looks at data differently and has multiple parameters governing how it works. It is difficult to know what the right method is before starting an analysis. These decisions are difficult for experts and can be extremely intimidating for non-expert users such as most biologists or clinicians.

A Path to Broader Applications

How can machine learning be democratized given its complexities? The good news is that a new discipline, automated machine learning (AutoML), has grown organically over the last 10 years, leading to the development of algorithms capable of making many of the decisions outlined above and yielding an optimal set of methods and hyperparameters for a given data set and problem[14,15]. Methods and software such as AutoSklearn, AutoWeka, and the Tree-based Pipeline Optimization Tool (TPOT) were among the first automated approaches to be developed and released as open-source[16-18]. AutoML methods were later developed and released by major

technology companies such as Google and Amazon. These methods are maturing quickly and becoming an integral part of the machine learning toolbox. As an indication of how new the field is, the first AutoML books were published in 2018-19[19-21] and the first conference **1st International Conference on Automated Machine Learning**, was held in 2022.

We believe that AutoML will greatly accelerate cancer research and care by bringing machine learning technology to everyone who wants to use it. To make this a reality, bioinformaticians and data scientists will need to work with information technology teams to make these computational resources available to cancer institutes and their investigators and to support them. We have begun such training and implementation at our cancer center. Successful deployment of AutoML tools and their use by the cancer center research community will require far less effort than that to support the hundreds of algorithms and software packages necessary to piece together a machine learning pipeline and the intensive consultation with users that would require.

REFERENCES

1. Daniel, M., Hajek, P. & Nguyen, P. H. CADIAG-2 and MYCIN-like systems. *Artif Intell Med* 9, 241-259 (1997). #

2. Cruz, G. P. & Beliakov, G. On the interpretation of certainty factors in expert systems. *Artif Intell Med* 8, 1-14 (1996). #

3. Bartels, P. H., Thompson, D. & Weber, J. E. Expert systems in histopathology. V. DS theory, certainty factors and possibility theory. *Anal Quant Cytol Histol* 14, 165-174 (1992).

4. Yu, V. L. *et al.* Antimicrobial selection by a computer. A blinded evaluation by infectious diseases experts. *JAMA* 242, 1279-1282 (1979).

5. van de Sande, D. *et al.* Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Health Care Inform* 29 (2022). #

6. Francesconi, E. The winter, the summer and the summer dream of artificial intelligence in law: Presidential address to the 18th International Conference on Artificial Intelligence and Law. *Artif Intell Law (Dordr)* 30, 147-161 (2022). #

7. Crevier, D. *AI : the tumultuous history of the search for artificial intelligence*. (Basic Books, 1993).

8. Rachlin, H. Making IBM's Computer, Watson, Human. *Behav Anal* 35, 1-16 (2012). #

9. Ferrucci, D. *et al.* Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31, 59 (2010). #

10. Harish, V., Morgado, F., Stern, A. D. & Das, S. Artificial Intelligence and Clinical Decision Making: The New Nature of Medical Uncertainty. *Acad Med* 96, 31-36 (2021). #

11. Kandoi, G., Acencio, M. L. & Lemke, N. Prediction of Druggable Proteins Using Machine Learning and Systems Biology: A Mini-Review. *Front Physiol* 6, 366 (2015). #

12. Koprowski, R. & Foster, K. R. Machine learning and medicine: book review and commentary. *Biomed Eng Online* 17, 17 (2018). #

13. Rodriguez-Barranco, M., Tobias, A., Redondo, D., Molina-Portillo, E. & Sanchez, M. J. Standardizing effect size from linear regression models with log-transformed variables for meta-analysis. *BMC Med Res Methodol* 17, 44 (2017). #

14. Waring, J., Lindvall, C. & Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med* 104, 101822 (2020). #

15. Manduchi, E., Romano, J. D. & Moore, J. H. The promise of automated machine learning for the genetic analysis of complex traits. *Hum Genet* (2021). #

16. Nantasenamat, C. *et al.* AutoWeka: toward an automated data mining software for QSAR and QSPR studies. *Methods Mol Biol* 1260, 119-147 (2015). **#**

17. Wang, G. *et al.* Rapid identification of human ovarian cancer in second harmonic generation images using radiomics feature analyses and tree-based pipeline optimization tool. *J Biophotonics* 13, e202000050 (2020). **#**

18. Olson, R. S., Bartley, N., Urbanowicz, R. J. & Moore, J. H. in *GECCO '16: Proceedings of the Genetic and Evolutionary Computation Conference 2016.* (ACM Digital Library).

19. Das, S. & Cakmak, U. *Hands-On Automated Machine Learning*. 1st edition edn, (Packt Publishing, 2018).

20. Mukunthu, D., Shah, P. & Tok, W.-H. *Practical automated machine learning on Azure : using Azure machine learning to quickly build AI solutions*. First edition. edn, (O'Reilly, 2019).

21. in *Automated Machine Learning The Springer Series on Challenges in Machine Learning* (ed Lars Kotthoff Frank Hutter, Joaquin Vanschoren) (Springer Cham, 2019).

Our Mission

The Association of American Cancer Institutes (AACI) represents 104 premier academic and freestanding cancer centers in the United States and Canada. AACI is accelerating progress against cancer by empowering North America's leading cancer centers in their shared mission to alleviate suffering.

About AACI *Commentary*

To promote the work of its members, AACI publishes *Commentary*, a monthly editorial series focusing on major issues of common interest to North American cancer centers, authored by cancer center leaders and subject matter experts.